多维特征下社会化问答社区答案排序研究*

■ 易明 张婷婷 李梓奇
华中师范大学信息管理学院 武汉 430079

摘 要：[目的/意义]研究多维特征对社会化问答社区答案排序的影响，以提高问答社区服务质量并尽可能优化用户体验。[方法/过程]从答案特征、回答者特征和投票者特征多个维度构建社会化问答社区答案排序特征体系，比较基于深度学习、树、神经网络、支持向量机等11种排序学习算法在问答社区数据集上的适用性，并训练随机森林分类算法，得到每个特征的重要程度。[结果/结论]实验结果表明，基于深度学习的排序学习算法在NDCG@k和MRR指标上的性能均优于其他排序算法，投票者的影响力特征最为重要，其次是答案内容特征，最后是回答者的专业度特征，可以考虑从增加答案排序方式的多样性和提高答案排序算法的综合性两个维度进一步优化答案排序。

关键词：社会化问答社区 答案质量 排序学习算法 深度学习算法
分类号：G250
DOI：10.13266/j.issn.0252-3116.2020.17.011

1 引言

Web2.0时代背景下，社会化问答社区（social Q&A communities）已成为用户知识获取和互动交流的重要网络社区。社会化问答社区经历了十多年的发展，日渐完善，同时深刻地影响着用户的知识获取、社交行为等方方面面[1]，但是它在给用户带来互动便利的同时，“知识泛滥”“知识超载”等问题也随之出现。社会化问答社区的答案排序研究有利于提供高质量问题答案、提升服务质量，现已成为社会化问答社区实践和研究领域的关注重点。目前社会化问答社区答案排序研究主要从答案结构特征、答案文本特征或回答者特征3个维度展开，特征维度选取并不全面。在答案排序算法研究中，多数研究集中在排序算法的优化或提出上，导致现有排序算法的种类繁多，因此缺乏评估不同类别排序算法在社会化问答社区数据集中的适用性的研究。同时，从现实需求方面考虑，目前社会化问答社区用户群体基数大，用户的信息需求存在极大差异，现有的答案排序方式很难满足不同用户的需求，如相比于用户投票排序结果，有的用户想看到领域专家的回答，有的用户想看到内容丰富的回答，而有的用户可能更想看到主观/客观性较强的回答等。鉴于此，本文将针对上述社会化答案排序理论研究与实践应用的局限性，通过构建多维度答案排序特征体系，将多维排序特征体系融入排序算法，以解决特征维度不全面以及不同算法适用性的问题，提高问答社区服务质量以优化用户体验。

2 社会化问答社区答案排序研究现状

2.1 社会化问答社区答案排序特征指标

综观国内外研究现状，可以发现学者大多利用答案特征以及回答者特征对答案进行排序研究，也有少部分学者从投票人群特征的角度进行研究。

2.1.1 答案特征

答案特征主要包含答案外部特征和答案内部特征。

答案外部特征指可以直接统计得出的特征。通过对文献进行梳理，发现研究者常用的答案外部特征包括：①答案长度[2-4]；②链接、图片代码等的数量，答案包含此类内容越多，则表明答案的内容越丰富[2-5]；

* 本文系国家社会科学基金项目“基于人类动力学的信息网络信息交流行为研究”（项目编号：16BTQ076）和中央高校基本科研业务费重大培育项目“智慧图书馆系统关键技术与应用研究”（项目编号CCNU18JCXK04）研究成果之一。
作者简介：易明（ORCID：0000-0002-4864-6025），教授，博士生导师；张婷婷（ORCID：0000-0002-5068-8232），硕士研究生；李梓奇（ORCID：0000-0003-1880-2426），博士研究生，通讯作者，E-mail：lzq911015@qq.com.
收稿日期：2020-01-13 修回日期：2020-04-09 本文起止页码：103-113 本文责任编辑：易飞

③点赞数、评论数、反对数、浏览数等，对答案进行点赞、评论等行为属于浏览用户的行为，参与到该答案的人数越多，则该答案越流行，越有可能是高质量的答案[5]；④名词、动词、疑问词等的数量，具有良好结构以及包含合理数量的浅层句法特征的答案更有可能是好的答案[2-3,6]；⑤答案数量，包括句子数量等[7]；⑥其他特征，如答案问题比率[2]、相同词语序列[6,8]；等等。

答案内部特征主要指蕴含在文本中，无法直观表现出来的特征。通过梳理发现学者所使用的答案排序内容特征大致包含三类：①问题答案相似度。问题与答案的主题相似度越高，则该答案越有可能是高质量答案[2,4-5,7]；②答案相似度。计算同一问题下不同答案的相似度可以过滤掉无关答案[9-10]；③情感极性。具有正向情感倾向的答案更有可能是一个好的答案[1-2]。

2.1.2 回答者特征

回答者特征维度主要通过对回答者专业程度的衡量来表示，通过对现有文献的梳理，可以发现衡量用户在某一话题或问题领域的专业程度，不同的研究者有不同的表述方式，如用户权威度、用户专业度、用户专业权威度等[3-5,8,11-14]，本文统一将其称为用户专业度。研究者利用不同的指标来衡量用户专业度，比如通过用户披露的个人擅长领域、关注领域、关注关键词等特征来衡量用户对某一领域的专业程度[12]；用户在某一主题下回答的问题数越多，提问的问题越少，则用户在这个主题下的专业度越高[5]；利用用户在某一领域的标准化最佳答案数量和答案的准确率来衡量用户的专业度[4]等；此外，不同类型的问题对答案质量也存在一定的影响，主要是通过问题分类模型研究改善分类效果，提升信息保存价值等[15]；答案的回答时间与答案质量之间有一定的关系，用户生产高质量的答案需要花费更多时间，使用移动端回答速度更快，但是质量更低，匿名者回答速度相比不匿名者速度也更快，但是很难产生高质量答案[16]。

2.1.3 投票者特征

S. Geerthik 等[17]从投票者特征的角度出发，构建新颖有效的答案排序模型，用到的特征指标包括：回答者的粉丝数、来自粉丝的点赞数量、来自非粉丝的点赞数量、专家点赞数、专家反对数、来自该问题下其他回答者的点赞数、非粉丝的反对数。崔宇佳等[18]利用基于特征的 Borda Count 排序投票法对多评价标准的结果进行融合。

2.2 社会化问答社区答案排序方法

2.2.1 利用现有的答案排序模型

一些研究学者将相关理论研究思想与现有的答案排序模型相结合，达到优化排序模型的效果。M. Surdeanu 等[6]把问题与答案间的相似度，提问术语在答案中出现的密度、频率到网络相关特征结合到感知排序模型和 SVMRank 模型中，从而优化模型。原立伟[11]借助迁移学习的思想对传统的排序学习方法进行改进，Ranking SVM 在 P@ N、MAP、NDCG 等指标上表现更优。田作辉[4]提出了基于质量检测和排序的答案选取方法，过滤掉低质量的答案后再对答案进行排序在准确率方面要优于直接对答案排序。Z. M. Zhou 等[8]将问答社区用户概况信息融入到 SVMRank、ListNet 排序模型，排序特征集中加入了用户相关特征后，在 MRR、P@ N 指标性能方面对答案排名更有效。

2.2.2 构建新的答案排序模型

一些研究学者提出新的答案排序模型。H. Toba 等[2]提出了一种混合层次分类模型，该模型预先设定 6 类问题，计算每个问题属于不同类别的概率，然后在 6 个子问题分类模型下计算答案质量，最后结合问题和答案分类模型综合评定答案质量好坏，该框架对于识别高质量答案是有效的，相比较于其他模型准确度更高。Y. Shen 等[19]提出了一种新的体系结构，利用包含词汇和顺序信息的相似矩阵，将信息放入深层体系中寻找潜在合适的答案，该方法在提高问答匹配精度方面具有一定的潜力，在 DCG@ P 评价指标方面优于基准算法。袁健等[5]提出了一种基于混合式的社区问答答案质量评价模型，该方法可有效地对答案进行质量评分，在 NDCG@ P 评价指标方面优于 PLSA 和 TSPR 模型。Z. Zhao 等[20]提出了新的基于 RNN 的异构非对称排序学习模型，该算法在 NDCG、P@ N、Accuracy 方面体现了优于其他先进算法的性能。

2.2.3 构建回答者排序模型

还有一些研究者提出先对回答者进行排序找到专家，然后直接将专家的答案作为最佳答案。X. Liu 等[13]提出了 ZhiHuRank 算法，根据问题和专家领域的链接结构和主题相似性，确定了用户权威性的排名，该算法在 MRR、NDCG 性能方面优于其他算法。L. Yang 等[21]提出了 CQARank 模型，CQARank 不仅可以找到具有相似主题兴趣的专家，而且可以根据社区中的问答投票历史找到具有高度专业知识的专家，该模型在 MRR、P@ N、CDR@ P 指标方面优于其他算法。刘瑜等[11]通过分析用户的行为，进一步提出 RTEM 模型

易明,张婷婷,李梓奇.多维特征下社会化问答社区答案排序研究[J].图书情报工作,2020,64(17):103-113.

(Related Topic Expertise Model),RTEM 模型与 TEM 模型相比,NDCG、Spearman、Kendall 的效果更优。

2.3 研究述评

综上所述,现有对社会化问答社区答案排序研究的成果较多,如从答案结构特征、答案文本特征或回答者特征维度选择答案排序特征,利用已有经典排序算法或提出新的排序算法对问答社区答案进行排序,或从回答者角度通过识别领域专家的方式对答案进行排序研究。

但是目前研究依然存在不足:①对答案排序特征的选择主要集中在答案特征和回答者特征两方面,而考虑投票者特征的研究相对较少。②多数研究侧重于答案排序算法的研究,而对不同类别的排序学习算法进行对比分析的研究相对较少。针对这两点,本文从多维特征角度试图构建答案排序特征体系,并比较不同排序算法在评价指标上的性能差异和分析不同排序特征对于答案排序结果的贡献度,最后基于本文的研究提出社会化问答社区答案排序优化策略。

3 社会化问答社区多维答案排序特征体系构建

3.1 社会化问答社区多维答案特征

答案排序的目的是为用户提供高质量的回答,答案本身是衡量答案质量的直接因素。其次,答案来源的可靠性也间接影响了答案质量。此外,社会化问答社区中最核心的要素是用户,大量用户不断地在平台获取或分享信息,社会化问答社区为了实现用户间的问答信息交流,引入一些运营机制,其中包括投票机制,用户可以根据答案质量选择"赞"或"踩"的投票操作,平台会根据用户的投票情况对回答进行排序,投票者对答案的评价在一定程度上也表明了答案质量的高低。因此,本文从答案特征、回答者特征以及投票者特征 3 方面构建答案排序的特征集合。

3.1.1 答案特征

答案特征最重要的是答案质量,本文从以下 4 个方面对答案质量进行测量:①答案长度(length)。回答越长表明了答案所含信息量越大,即答案长度与答案所包含的信息成正比[22],同时答案长度也反映了回答者的努力程度[23]。②答案与问题的相似度(similarity)。从语义层面反映了答案的质量,答案与问题主题间的相似度越高,说明答案与问题的语义越接近,答案更有可能解决提问者的信息需求。与问题相似度较低的答案,通常是无用的回答,应该被过滤掉[5]。③答案信息熵(entropy)。信息熵可以反映答案的多样性,答案信息熵值越大,表示答案内容越丰富。相比于熵值低的答案,高熵值的答案质量更高[3]。④答案种类(form)。答案中文字、图片、链接等内容种类的个数。外部链接、图片或图表,可用于证明回答者观点的正确性和可信度,此外,图片或图表将复杂的语言论述简化,便于用户理解。因此,答案种类丰富的回答更可能是高质量答案。

3.1.2 回答者特征

回答者特征主要包括以下两方面:

(1)回答者对所答问题的专业程度。回答者的专业程度越高,说明回答者越具备回答该问题的专业知识和特长,其答案越有可能是高质量回答[24]。本文从回答者提问、回答的数量和质量来衡量回答者的专业度:①回答者历史相似问题的回答专业度(a_aSpecialty)。回答者历史回答过的相似问题越多,答案的点赞数越多,说明回答者越有可能是该问题领域的专家。通常,专家的答案相比于普通用户的回答质量更高[5]。②回答者历史提问相似问题的专业性(a_qSpecialty)。回答者历史提问过相似问题,说明回答者在回答该提问之前,积累过相关知识,具有一定的发言权。

(2)回答者在社区的影响力。用户的影响力来自于其可信任度或专业能力[25],因此,高社区影响力的回答者其答案质量可能更高。通常情况下,用户参与社区的时间长短、提问和回答的数量及质量、赞同数、粉丝数量等指标都能体现用户的影响力,本文主要从以下几方面进行测量:①回答者所获得的点赞数(a_voteup)。该特征反映了回答者的社区贡献以及所获得的成就。②回答者被关注数量(a_following)。从社会网络视角分析,被关注数反映了个体在社会网络中所拥有的社会资本。被关注数越多,说明回答者的影响力越大。③回答者的回答数量(a_aNum)。④回答者的提问数量(a_qNum)。这两个指标也反映了回答者的社区参与度和贡献。

3.1.3 投票者特征

大多数问答社区以大众投票对答案进行排序,投票者对答案的评判(点赞或反对)可以反映答案质量。S. Geerthik 等[17]认为,若答案 A 中意见领袖的点赞数高,那么答案 A 应该排在答案 B 前面;同样地,若答案 A 中意见领袖的反对数较高,那么答案 A 应该排在答案 B 后面。同时在社会化问答社区中,投票者对答案也起到了再传播的作用,投票者对答案的态度可能影响其他用户对答案的认知。因此,投票者特征影响了答案排序。投票者包括点赞者和反对者两类用户,某一答案点赞者的权威性越高,反对者的权威性越低,那

么该回答越应排名靠前,反之,排名应靠后。为了维护社区和谐氛围,问答社区不显示答案的反对票数以及反对者信息,我们无法获取反对者数据。因此,本文仅考虑点赞者特征对答案排序的影响。本文从以下两个方面对点赞者的权威性进行测量:点赞者平均所获点赞数(v_voteup)、点赞者平均粉丝数(v_following)。

本文涉及到的答案排序特征及其计算方法如表 1 所示:

表 1 答案排序特征及计算方法

特征类型	特征名称	英文缩写	计算方法
答案特征	答案长度	length	答案包含汉字、英文词、数字数量
	答案与问题主题相似度	similarity	主题模型 + 余弦相似度
	答案信息熵	entropy	信息熵公式
	答案种类	form	答案文字、图片、链接等内容种类的个数
回答者特征	回答者历史相似问题的回答专业度	a_aSpecialty	$\sum_{i=1}^{n} s_i v_i$
	回答者历史提问相似问题的专业性	a_qSpecialty	$\sum_{i=1}^{n} s_i a_i$
	回答者所获得的点赞数	a_voteup	回答者累积获得的点赞数
	回答者被关注数量	a_following	回答者被其他用户关注的数量
	回答者的回答数量	a_aNum	回答者历史回答问题数量
	回答者的提问数量	a_qNum	回答者历史提问题的数量
投票者特征	点赞者平均所获点赞数	v_voteup	答案下点赞者所获点赞数的均值
	点赞者平均粉丝数	v_following	答案下点赞者粉丝数的均值

3.2 知乎社区多维答案特征体系构建分析

本研究的实验对象为知乎,爬虫爬取了知乎的问题及回答界面、用户界面,共随机爬取 1 976 个问题、108 211 条答案以及 71 280 个回答者信息和 2 632 660 个投票者信息。因有的用户账号被知乎社区停用,因此其个人信息无法获取,故删除此种无效数据,因此,本文实际用于实验分析的数据共 95 021 条答案及相关数据。本部分首先对数据进行特征计算,根据答案特征值结果的相关系数,合并相关系数较高的变量并删除一部分相关系数低的变量,从而最终得到答案排序指标体系。

(1)按照表 1 答案排序特征的计算方法进行特征计算,可得表 2。

表 2 答案特征值

变量	值	变量	值
length	143	form	1
similarity	0.57	a_aSpecialty	26.587
entropy	5.41	a_qSpecialty	20.208
a_voteup	172	a_qNum	19
a_following	134	v_voteup	136
a_aNum	84	v_following	16

(2)本文采用最小值 - 最大值标准化的方法,将实验数据值映射到[0,1]区间。标准化以后的实验数据是[0,1]之间的连续正态分布变量,因此本文选择皮尔森相关数法计算各变量之间的相关系数,构建答案排序特征相关系数矩阵,如表 3 所示:

表 3 答案排序特征相关系数矩阵

	length	similarity	entropy	form	a_aSpecialty	a_qSpecialty	a_voteup	a_following	a_aNum	a_qNum	v_voteup	v_following
length	1	-0.02	**0.68**	0.21	0.06	0.04	0.06	0.03	0.01	0.03	0.07	0.11
similarity		1	0.07	-0.01	0	-0.01	0	-0.01	0	-0.01	0	0
entropy			1	0.15	-0.02	-0.03	-0.02	-0.06	-0.08	-0.05	-0.02	0.02
form				1	0.06	0.04	0.06	0.04	0.02	0.03	0.07	0.09
a_aSpecialty					1	0.36	**0.96**	0.32	**0.6**	0.34	0.11	0.17
a_qSpecialty						1	0.37	0.31	0.39	**0.67**	0.11	0.14
a_voteup							1	0.33	**0.61**	0.35	0.12	0.18
a_following								1	0.4	0.37	0.14	0.18
a_aNum									1	**0.48**	0.11	0.18
a_qNum										1	0.11	0.15
v_voteup											1	0.3
v_following												1

（3）筛选变量，构建指标体系。Pearson 相关系数的绝对值越大，变量的相关关系越强，绝对值越小，变量的相关关系越弱，其中，0.8－1 表明极强相关，0.6－0.8 为强相关，0.4－0.6 为中等程度相关，0.2－0.4 为弱相关，0－0.2 为极弱相关或者无相关。本文综合考虑变量的数量与相关系数的数值，认为应取 0.4 为阈值，相关系数大于 0.4 为相关，小于0.4 为不相关。答案信息熵（entropy）和答案长度（length）相关系数为 0.68，表明两者呈正相关，答案信息熵（entropy）与答案长度（length）都能衡量答案的信息内容，答案信息熵是量化答案信息的重要概念，相比于答案长度，更能体现答案信息的信息质量，根据特征对信息量化模型的拟合优度贡献程度舍弃变量 length。回答者历史相似问题的回答专业度（a_aSpecialty）与回答者历史回答数（a_aNum）的相关系数为 0.60、与回答者所获得的总点赞数（a_voteup）的相关系数为 0.96，回答者历史相似问题的回答专业度（a_aSpecialty）的数据来源是回答者的历史回答的问题集合 Q 以及历史回答所获得的点赞数集合 V，而回答者的历史回答的问题集合 Q 中包含回答者的回答数量（a_aNum），历史回答所获得的点赞数集合 V 中包含回答者所获得的点赞数（a_voteup），因此保留综合指标 a_aSpecialty。同理，回答者历史提问相似问题的专业性（a_qSpecialty）和回答者历史提问数（a_ qNum）相关系数为0.67，保留综合指标变量 a_qSpecialty。删除特征 length、a_aNum、a_voteup、a_qNum 后，其他特征的相关系数均小于 0.4。

本文选择答案与问题的相似度、答案信息熵、答案种类、回答者历史相似问题的回答专业度、回答者历史提问相似问题的专业性、回答者的粉丝数、点赞者平均所获点赞数、点赞者平均粉丝数共 8 个特征作为最终的特征集。本文所构建的答案排序特征体系如图 1 所示，与已有答案排序特征体系相比，本文综合考虑了答案、回答者以及投票者的特征。已有研究选取答案和回答者特征构建指标体系，或仅从投票者维度选取指标构建答案排序模型，未从这 3 个维度综合选取排序特征。其次，本文从理论研究的角度选取社会化问答社区答案排序特征集后，采用 Pearson 相关系数的方式选择特征，构建稳定的答案排序特征体系。而大多数已有研究缺少对特征进行选择的环节，会影响排序算法的性能。

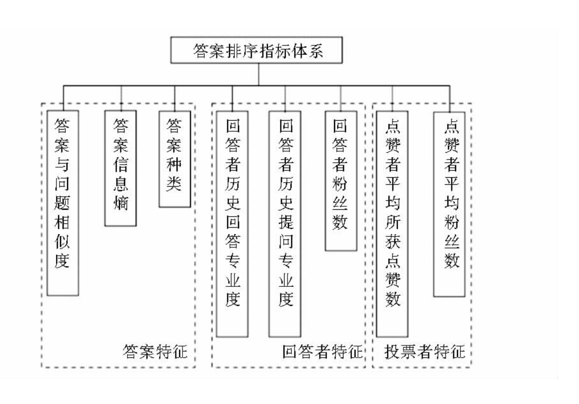


图 1　答案排序特征体系

4　基于多维特征的社会化问答社区答案排序实验分析

本实验的研究任务有两个，一是比较不同算法的性能，二是比较不同答案排序特征的重要性，为社会化问答社区答案排序任务评估排序特征和排序方法。本实验的大致流程如图 2 所示：

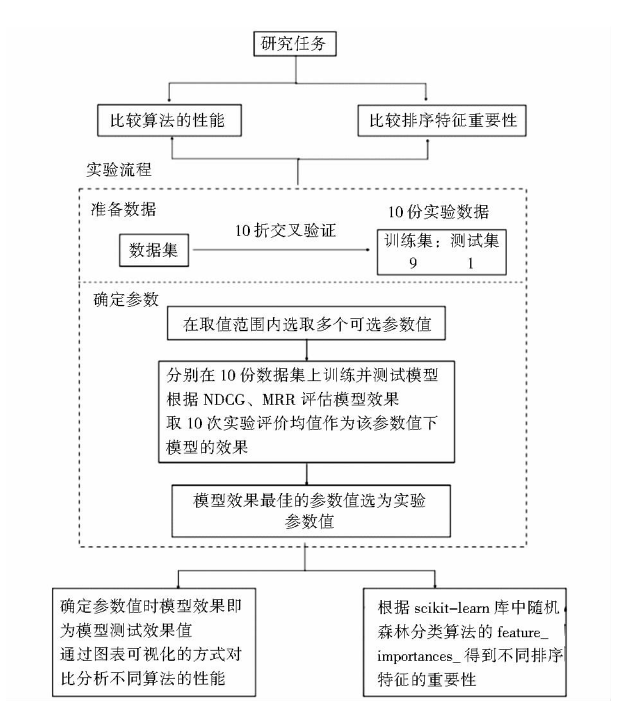


图 2　社会化问答社区答案排序实验设计

4.1　基于多维特征的社会化问答社区答案排序实验过程

4.1.1　实验工具

现有排序学习数据相关性标注策略通常指定文档是否与查询相关（如二进制判断 1 或 0），或进一步指定相关程度（如多种有序类别，Perfect，Excellent，Good，Fair，Bad）[26]。该相关性判定多采用人工进行标注，人

工标注结果质量高,但成本较大。T. Joachims 等[27-28]提出一种巧妙的方法,利用搜索引擎的点击数据获得文档相关性的标注。本文借鉴该研究思想,利用答案的点赞数作为相关性判断的依据。社会化问答社区设置了投票机制,用户可以点赞对自己有用的答案,也可以点踩对自己无用的答案。用户投票类似于同行评审机制,用户在投票的过程中虽存在一定的主观性,但已被证明大众参与的投票模式在一定程度上是有效的[29]。由于无法爬取到知乎社区答案反对数据,因此,本文仅根据答案点赞数对答案相关性进行标注,将答案点赞数大于平均点赞数的标记为 1,将答案点赞数小于平均点赞数的标记为 0,这种标注方式相对粗放,但是在本研究中,由于数据中答案的区分度相对较低,采用细粒度的标注会为后面的研究带来更大的误差,考虑到标注的粒度不是研究的重点,因此研究在标注时采取粗粒度的区分方式。根据该方法,本实验将 30 261 条实验数据标记为 1,64 760 条实验数据标记为 0。

本文采用的 11 种排序算法及其开源工具如表 4 所示:

表 4 本研究使用的排序算法与工具

序号	算法	开源工具	序号	算法	开源工具
1	Ranking SVM	SVM Rank	7	MART	RankLib
2	RankNet	RankLib	8	LambdaMART	RankLib
3	ListNet	RankLib	9	Linear Regression	RankLib
4	RankBoost	RankLib	10	Coordinate Ascent	RankLib
5	AdaRank	RankLib	11	Deep Learning	TF-Ranking
6	Random Forest	RankLib			

4.1.2 评价指标

本文采用 NDCG@ k 和 MRR 两种指标衡量各个排序算法性能。NDCG 考虑文档的多级相关度,根据文档在结果列表中的位置测量文档的有用性,是衡量排名质量的重要指标。k 表示对于前 k 位的排序结果计算 NDCG 值,本文 k 分别取值 1,3,5 和 10。MRR 则只关注排序结果中第一个文档的相关性,用户通常从上往下浏览答案,当找到合适的答案时就结束本次搜索,NDCG@ k 是最常用的衡量排序结果的指标,主要是评价答案排序列表的质量好坏,NDCG 有两个准则,一是相关程度高的结果对 NDCG 的影响更大,二是相关程度高的结果越靠前,NDCG 的值越高,另外,NDCG 会考量 k 个答案组成的排序列表与理想答案排序列表的差距,因此 NDCG 主要是对 k 个答案所组成的排序列表的整体评价,而 MRR 是相对简单的一种评价指标,主要衡量的是最准确的答案在列表中的位置,其位置越靠前,MRR 值越高。可以看出,NDCG@ k 与 MRR 虽然都是对答案排序的评价,但是关注的对象不同,前者更关心整个答案列表,而后者更关心最准确的答案。本文采用 NDCG@ k 和 MRR 的目的是通过不同的评价指标来对算法进行衡量。

4.2 以深度学习排序算法为例的实验分析

本文以深度学习排序算法为例演示实验流程。实验流程包含 3 个部分:准备数据集、确定实验参数值、对比不同排序算法在评价指标上的排序表现。算法参数值的选取很重要,参数值选择不当会导致模型出现过拟合和欠拟合的情况,进而影响到算法的排序性能。本实验主要演示深度学习排序算法的参数值选择的过程。

本文将 num_features 设置为了实验数据集中每个答案的特征数 8,然后重点对 num_train_steps(训练步数)参数进行了调节,其余参数均保持默认值不变。首先,本文先将 num_train_steps 设置为 20 000,算法训练过程中评价指标的变化如图 3 所示。可以发现当 num_train_steps 大致在[6 000,10 000]的取值范围内,评价指标值处于最优状态且趋势较为平稳,当 num_train_steps 大于 10 000 时,算法的性能变差。因此,本文分别将 num_train_steps 设置为 6 000、7 000、8 000、9 000、10 000 进行实验,根据评价指标的大小,选择最优的参数估计值。

以 num_train_steps = 7 000 为例,对深度学习排序算法进行 10 折交叉验证。分别得到 10 次训练 7 000 次的评价指标值并计算这 10 次实验结果的平均值,作为 num_train_steps = 7 000 时深度学习排序算法的性能评价,结果见表 5。

根据以上实验步骤,num_train_steps 为不同值时,深度学习排序算法的性能评价结果如表 6 所示。综合比较,当 num_train_steps 取 8 000 时,深度学习排序算法的性能最好,因此,本文将 num_train_steps 的实验值定为 8 000。

4.3 实验结果分析

4.3.1 模型效果分析

分别对排序学习算法进行 10 折交叉验证,确定了算法的参数值,并以 10 次实验结果的平均值作为排序算法的评价结果,实验结果见表 7。

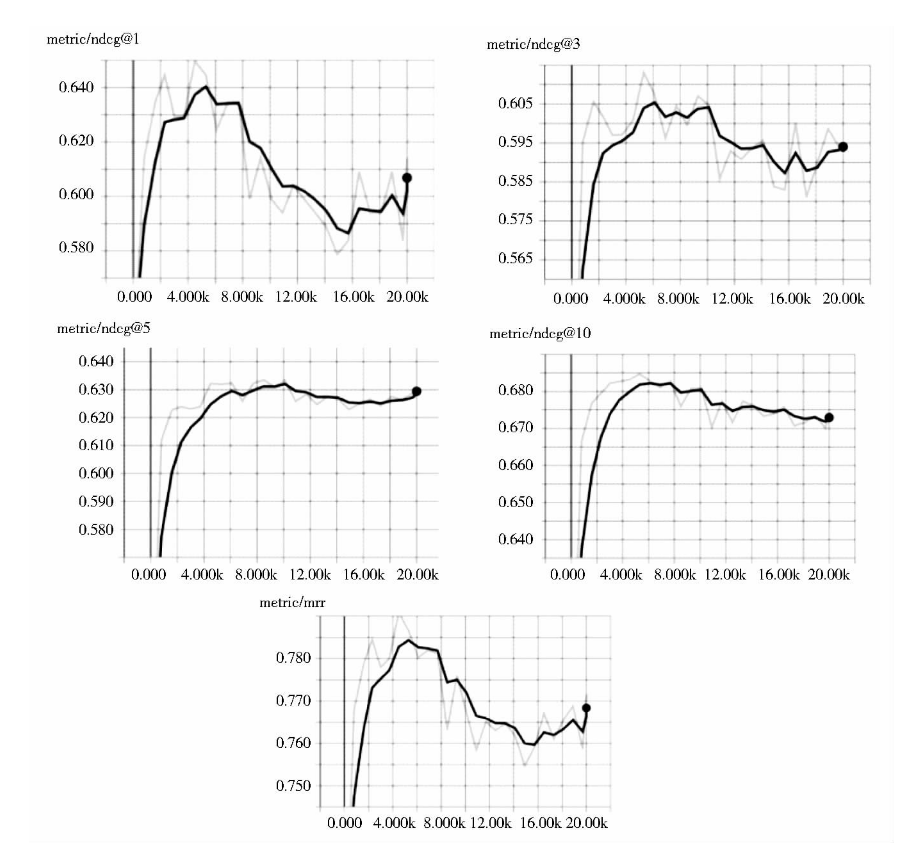


图 3　深度学习排序算法训练过程中评价指标的变化

表 5　num_train_steps = 7 000 时算法指标评价详情数据

序号	NDCG@1	NDCG@3	NDCG@5	NDCG@10	MRR
1	0.644 7	0.608 1	0.632 4	0.683 7	0.784 3
2	0.614 2	0.603 6	0.628 6	0.683 5	0.769 6
3	0.588 8	0.582 2	0.586 9	0.645 1	0.756 5
4	0.705 6	0.622 5	0.642 7	0.692 2	0.819 5
5	0.614 2	0.616 3	0.628 4	0.688	0.774 9
6	0.680 2	0.622 9	0.627 5	0.681 8	0.804
7	0.629 4	0.600 4	0.621 2	0.683 2	0.774
8	0.710 7	0.652 3	0.660 5	0.694	0.827 2
9	0.685 3	0.634	0.640 3	0.677 3	0.810 1
10	0.741 1	0.641 7	0.633 7	0.658	0.844 2
均值	0.661 4	0.618 4	0.630 2	0.678 7	0.796 4

表 6　num_train_steps 取不同值时算法性能比较

num_train_steps 取值	NDCG@1	NDCG@3	NDCG@5	NDCG@10	MRR
6 000	0.663 5	0.618 1	0.630 2	0.678 7	0.796 9
7 000	0.661 4	0.618 4	0.630 2	0.678 7	0.796 4
8 000	**0.665 0**	**0.624 3**	**0.634 1**	**0.679 9**	**0.799 5**
9 000	0.645 2	0.612 6	0.624 4	0.674 9	0.786 7
10 000	0.659 9	0.619 1	0.629 6	0.680 4	0.797 1

表 7　排序学习算法性能比较

排序学习算法	NDCG@1	NDCG@3	NDCG@5	NDCG@10	MRR
Ranking SVM	0.542 1	0.500 1	0.509 0	0.558 9	0.710 1
RankNet	0.574 6	0.526 4	0.520 9	0.570 7	0.726 5
ListNet	0.507 1	0.483 9	0.505 9	0.564 3	0.688 8
RankBoost	0.583 3	0.550 2	0.557 3	0.603 0	0.742 8
AdaRank	0.325 4	0.332 5	0.361 2	0.443 6	0.541 9
Random Forest	0.653 3	0.611 0	0.617 7	0.663 5	0.793 4
MART	0.653 3	0.611 0	0.616 8	0.662 7	0.793 1
LambdaMART	0.636 6	0.607 1	0.609 5	0.653 1	0.781 9
Linear Regression	0.416 2	0.408 2	0.444 7	0.525 3	0.620 8
Coordinate Ascent	0.560 9	0.517 6	0.518 2	0.567 0	0.723 7
Deep Learning	**0.665 0**	**0.624 3**	**0.634 1**	**0.679 9**	**0.799 5**

从输入数据样例(或损失函数)分类角度进行分析,同为神经网络排序学习算法的 RankNet 和 ListNet 相比较,基于 Pairwise 的 RankNet(学习率 lr = 0.000 05)算法在 NDCG@k 和 MRR 指标上,表现均优于基于 Listwise 的 ListNet(学习率 lr = 0.000 05)算法。同为树排序学习算法的 MART 和 LambdaMART 相比较,基于 Pairwise 方法的 MART(学习率 shrinkage or lr = 0.05)算法在评价指标上均略优于基于 Listwise 的 LambdaMART(学习率 shrinkage or lr = 0.01)算法。同样地,同为提升排序学习算法的 RankBoost 和 AdaRank

相比较,基于 Pairwise 方法的 RankBoost 算法在评价指标上均优于基于 Listwise 的 AdaRank(连续两轮学习之间的误差 tolerance = 0.002)算法。实验结果表明基于 Listwise 的排序方法不一定比基于 Pairwise 和 Pointwise 的排序方法好,其原因可能是 Listwise 排序方法虽然考虑了同一问题下的答案序列关系,但其很难找到合适的目标代替原有优化目标,也很难找到合适的优化算法求解目标[30]。

从机器学习技术分类角度进行分析,不同机器学习技术的排序算法在知乎问答社区数据集上表现性能各不相同。首先,基于深度学习的排序学习算法 Deep Learning(训练步数 num_train_steps = 8000)在 NDCG@k 和 MRR 指标上均优于其他算法。其次,基于树的排序学习算法 Random Forest(学习率 shrinkage or lr = 0.1)、MART(学习率 shrinkage or lr = 0.05)、LambdaMART(学习率 shrinkage or lr = 0.01)表现也要优于其他传统机器学习方法,其中 Random Forest 算法在三者之间性能更略胜一筹,可能原因是基于树的排序算法容易发生过拟合,而 Random Forest 是由多棵树组成,每棵树仅学习特征集的部分特征,最终分类结果由所有树投票决定,因此 Random Forest 可以在很大程度上减少过拟合。最后,基于提升的排序方法 RankBoost、基于支持向量机的 Ranking SVM(正则化参数 c = 0.01)、基于神经网络的排序方法 RankNet(学习率 lr = 0.000 05)以及基于梯度上升的排序方法 Coordinate Ascent(两种方案间的性能误差 tolerance = 0.001)的性能相对 AdaRank(连续两轮学习之间的误差 tolerance = 0.002)、ListNet(学习率 lr = 0.000 05)、Linear Regression(正则化参数 L2 = 1.0E-10)算法要好。

4.3.2 特征重要性分析

为了分析答案排序特征体系各个指标的重要性,本文选取随机森林算法进行特征重要性评估。随机森林算法的思想是判断每个指标特征在每个树节点上做了多大的贡献,然后取平均值并比较特征之间的贡献大小。作为分类器,随机森林使用“强变量”的一个小子集进行隐式特征选择,这使得它在高维数据上具有优异的性能[31]。随机森林的隐式特征选择的结果可以通过“基尼重要性”进行可视化[32],并且可以作为特征相关性的一般指标。这种特征重要性评分提供了特征相对排序的方法,并且在技术实现上是随机森林分类器训练的副产品:在随机森林的二叉树 T 中的每个节点 τ,使用基尼不纯度 $i(\tau)$(对熵的有效近似计算)寻求最佳分割,用于测量潜在分割在特定节点上将样本分为两类的程度。

$p_k = \frac{n_k}{n}$表示节点 τ 的样本总数 n 中,k 类样本数量为 n_k 的分数,其中 $k \in \{0,1\}$,那么,基尼不纯度表示为:

$$i(\tau) = 1 - p_1^2 - p_0^2 \qquad 式(1)$$

根据变量 θ 的阈值 t_θ,节点 τ 的样本被分割给两个子节点 τ_l 和 τ_r(具有各自的样本分数 $p_l = \frac{n_l}{n}$和 $p_r = \frac{n_r}{n}$,那么,节点 τ 的基尼不纯度减小 Δi,Δi 表示为:

$$\Delta i(\tau) = i(\tau) - p_l\, i(\tau_l) - p_r\, i(\tau_r) \qquad 式(2)$$

在对节点上所有可用变量 θ 的穷尽搜索中(随机森林将搜索限制在可用特征的随机子集[30]),并在所有可能的阈值 t_θ 上,$\{\theta, t_\theta\}$ 对可以确定最大 Δi。基尼不纯度的减小是由于森林中所有树 T 上的所有节点 τ 的最佳分割 $\Delta i_\theta(\tau, T)$ 的积累,对于每个变量 θ:

$$I_G(\theta) = \sum_T \sum_\tau \Delta i_\theta(\tau, T) \qquad (3)$$

基尼重要性指数 G 表明了选择特定特征 θ 进行分割的频率以及该特征对于分类问题的整体识别值有多大。

根据随机森林算法计算得到的特征重要性如图 4 所示:①投票者的特征. 点赞者所获得的点赞数均值、点赞者的粉丝数均值两个指标得分最高,说明投票者所具有的影响力越大,投票者对答案做出的评价越能反映答案的质量。②答案特征. 答案的信息熵和答案与问题的相似性两个指标得分较高,答案内容的丰富程度、答案与问题的语义相似度是客观反映答案质量的指标,答案内容越丰富、答案与问题的语义越相近,答案的质量越高,答案越应该排到前面。但答案形式的多样性重要性最低,说明相对于答案的外在表现形式,问答社区用户更加重视答案的内容。③回答者特征。可以发现回答者历史相似问题的回答专业度与回答者的影响力(粉丝数)两个指标比回答者历史提问相似问题的专业性指标更重要。因为回答者历史回答过的相似问题越多,答案的点赞数越多,说明回答者越有可能是该问题领域的专家。通常,专家的答案相比于普通用户的回答质量更高。而回答者历史提问相似问题的专业性指标仅能说明回答者历史提问过相似问题并获得过一些回答,但还是无法准确衡量用户是否掌握该话题领域的专业知识。

5 社会化问答社区答案排序优化策略

根据社会化问答社区答案排序研究结果,并结合

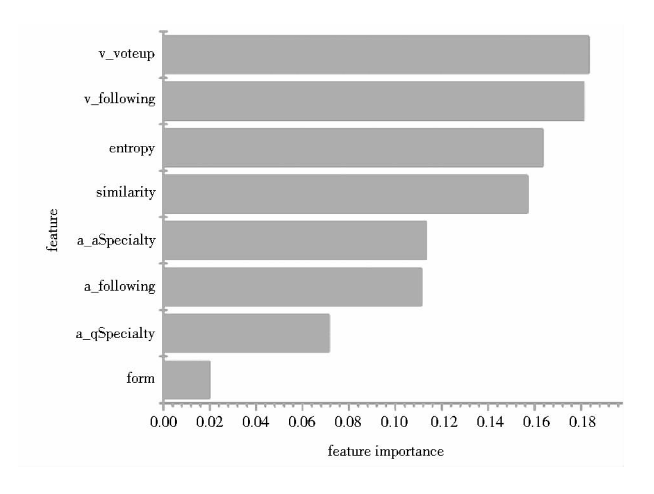


图 4　特征重要性对比分析

问答社区所面临答案排序无法有效满足用户需求的现实困境，本文从答案排序方式的多样性和答案排序算法的综合性提出答案排序优化策略。

5.1　增加答案排序方式的多样性

本文通过理论分析和数据检验最终确定了 8 个排序指标，并根据实验结果分析了 8 个排序指标的相对重要性。以本文的答案排序指标体系为例，为社会化问答社区提供多样排序选择的优化策略：

(1)根据信息熵或答案与问题的主题相似度特征对同一问题下的答案进行排序。答案与问题的相似度从语义层面反映了答案与问题的相似程度，而答案信息熵反映了答案的内容丰富度。将答案与问题的相似度高或答案信息熵大的答案排在前面可以让用户更快地看到更多内容丰富且与提问语义更相近的答案，而将与提问无关、抖机灵的答案排在了后面，节省了用户选择该类答案的时间成本。同时，增加了回答者(并非大 V 但认真回答问题的用户)的满足感和成就感，激励其持续分享知识的意愿，为社区平台增强了用户粘性。

(2)根据答案种类特征(或是否包含图片、超链接等特征)对同一问题下的答案进行排序。答案种类包括答案文字、图片、链接等内容。知乎问答社区中答案一般都很冗长，有的用户可能更希望看到简洁、清晰的论述，如使用图片或图表将复杂的语言论述简化。也有的用户可能更希望看到高可信度的答案，如答案中引用他人观点、文献等。

(3)根据回答者的领域专业度对同一问题下的答案进行排序。领域专家的回答也许答案并不长，也没有使用图片、超链接等表述方式，但其简短的回答通常可以解决提问者的困惑。对于以获取知识为目的的用户来说可能更倾向于浏览领域专家的回答，以便快速地获取专业知识，节省时间成本。

(4)根据投票者的影响力对同一问题下的答案进行排序。相较于普通用户对答案的点赞或点踩，用户可能更看重大 V(或领域专家)对答案的评价。

5.2　提高答案排序算法的综合性

本文从答案特征、回答者特征、投票者特征 3 个维度对答案排序特征进行了提取并分析了其对答案排序的重要程度。问答社区中除了本文爬取到的原始数据类型以外，还包含有大量的数据类型，如用户的点击流数据，用户访问时间、访问频次、访问地点以及海量的后端特征日志数据。社会化问答社区开发人员可以据此进行数据挖掘，提炼出丰富的答案排序特征集。此外，本文还比较了基于深度学习、树、支持向量机、提升等的排序算法在社会化问答社区数据集上的适用性。结果表明，基于深度学习的排序学习算法在 NDCG@ k 和 MRR 指标上的性能均优于其他排序算法。深度学习是近年来发展迅速的技术，已经成功应用于语音识别、图像识别、自然语言处理等多个领域。目前也有很多学者将卷积神经网络、递归神经网络等算法与排序学习研究相结合，并取得了可喜的研究成果。深度学习相较于传统的机器学习方法，有两大优点：

(1)适用于大样本数据量，且样本数据的增加能明显改善模型的结果。随着大数据时代的到来，问答网站掌握着越来越多的全量数据，传统的基于小样本的实验分析受到挑战，新的基于全样本的实证研究正在崛起。基于全样本的实证研究方法优势显著，样本量巨大，由于采用的是全样本数据，无需对实验数据进行有效性、内生性等检验，应用型大大增强，研究结果更接近于真实。

(2)传统的机器学习方法在进行模型训练之前，需要对原始数据进行处理和特征选择。而深度学习算法可以通过低维密集的特征，学习到以前没出现过的特征之间的关系，并且相比于线性模型大幅降低了对于特征工程的需求，因此，社会化问答社区开发人员在选择排序算法上可优先考虑深度学习排序算法。

6　结语

社会化问答社区现有的排序方法较为单一，最常使用的是根据用户点赞数和反对数对答案进行排序。这种大众评审机制在一定程度上能反映答案的质量，而一些抖机灵的回答或者段子往往因为有趣会获得很高的点赞数，但这些回答往往不能满足用户获取知识

的需求。因此,优化现有的排序算法是问答社区急需解决的问题。本文从两个角度为排序算法的优化提供了解决策略:①分析不同类别排序算法在社会化问答社区数据集上的排序性能;②分析答案排序特征的相对重要性,为社会化问答社区答案排序任务评估排序方法。本文根据实验结果以及实验分析,针对社会化问答社区现有的排序现状,从支持多种答案排序策略和构建基于深度学习的综合排序算法两个维度,为社会化问答社区答案排序优化提供建议。本文的研究仍存在一定缺陷,研究采用的数据集不充分,知乎社区没有公布答案获得的反对数量以及反对者信息,本文在构建答案排序特征阶段只考虑了点赞者的特征。本文的数据量也不充分,研究仅对知乎问答社区 10 万多条数据进行采集并进行实验研究,数据样本量相对于整个知乎来说太少。此外,研究缺乏对不同问题类别下答案排序异同的比较分析。而且,本文没有充分考虑点赞的不确定性,即点赞多的答案可能存在一定的从众效应,从而导致的话题性答案的点赞数相对于专业性答案的点赞数更多。今后将重点研究不同问题下,答案排序特征和答案排序模型的异同性以及点赞数与答案质量之间的关系。

参考文献:

[1] 李蕾,何大庆,章成志. 社会化问答研究综述[J]. 数据分析与知识发现,2018,2(7):1 – 12.
[2] TOBA H, MING Z Y, ADRIANI M, et al. Discovering high quality answers in community question answering archives using a hierarchy of classifiers [J]. Information sciences, 2014,47(8):101 – 115.
[3] 张鹏飞. 面向在线问答社区的问题检索与回答抽取技术研究与实现[D]. 长沙:国防科学技术大学,2015.
[4] 田作辉. 非事实类问题的回答选取[D]. 哈尔滨:哈尔滨工业大学,2013.
[5] 袁健, 刘瑜. 基于混合式的社区问答回答质量评价模型[J]. 计算机应用研究,2017,34(6):1708 – 1712.
[6] SURDEANU M, CIARAMITA M, ZARAGOZA H. Learning to rank answers to non-factoid questions from Web collections[J]. Computational linguistics, 2012, 37(2):351 – 383.
[7] 郭顺利,张向先,陶兴,等. 社会化问答社区用户生成答案质量自动化评价研究——以"知乎"为例[J]. 图书情报工作,2019, 63(11):118 – 130.
[8] ZHOU Z M, LAN M, NIU Z Y, et al. Exploiting user profile information for answer ranking in CQA[C]//21st World Wide Web conference 2012. Lyon:ACM Press, 2012:767 – 774.
[9] 程亚男, 王宇. 基于语义情感相似度的问答社区回答排序研究[J]. 情报科学, 2018,36(8):72 – 76,83.
[10] 廉鑫. 社区问答系统中若干关键问题研究[D]. 天津:南开大学,2014.
[11] 刘瑜, 袁健. 基于 RTEM 模型的问答社区候选回答排序方法[J]. 电子科技, 2016, 29(5):130 – 134.
[12] 原立伟. 社区问答系统中回答排序迁移学习的方法研究[D]. 昆明:昆明理工大学,2017.
[13] LIU X, YE S, LI X, et al. ZhihuRank: a topic-sensitive expert finding algorithm in community question answering Websites[C]//Advances in Web-based learning-ICWL 2015. Guangzhou: Springer International Publishing, 2015:165 – 173.
[14] LI B, KING I, LYU M R. Question routing in community question answering: putting category in its place[C]//ACM conference on information and knowledge management. Glasgow: ACM Press, 2011:2041 – 2044.
[15] 罗毅,曹倩. 基于 RIPA 方法的社会问答平台答案质量研究[J]. 图书情报工作,2015,59(3):126 – 133,25.
[16] 袁毅,杨莉. 问答社区用户生成资源行为及影响因素分析——以百度知道为例[J]. 图书情报工作,2017,61(22):20 – 26.
[17] GEERTHIK S, RAJIV G K, VENKATRAMAN S. Respond rank: improving ranking of answers in community question answering[J]. International journal of electrical & computer engineering, 2016,6(4):1889 – 1896.
[18] 崔宇佳,张一迪,王培志,等. 基于多评价标准融合的医疗数据特征选择算法[J]. 复旦学报(自然科学版),2019,58(2):250 – 255,268.
[19] SHEN Y, RONG W, SUN Z, et al. Question/answer matching for CQA system via combining lexical and sequential information [C]//29th AAAI conference on artificial intelligence. Austin: AAAI Press, 2015:275 – 281.
[20] ZHAO Z, LU H, ZHENG V W, et al. Community-based question answering via asymmetric multi-faceted ranking network learning [C]//Proceedings of the 31th AAAI conference on artificial intelligence. San Francisco:AAAI Press, 2017:3532 – 3539.
[21] YANG L, QIU M, GOTTIPATI S, et al. CQArank: jointly model topics and expertise in community question answering[C]//ACM international conference on information & knowledge management. San Francisco:ACM Press, 2013:99 – 108.
[22] JEON J, CROFT W B, LEE J H, et al. A framework to predict the quality of answers with non-textual features[C]//The 29th annual international ACM SIGIR conference on research and development in information retrieval. Washington: ACM Press, 2006: 228 – 235.
[23] 王乐. 社会化问答社区知识贡献和知识互动质量研究[D]. 哈尔滨:哈尔滨工业大学,2016.
[24] 王秀丽. 网络社区意见领袖影响机制研究——以社会化问答社区"知乎"为例[J]. 国际新闻界,2014,36(9):47 – 57.
[25] PERRY-SMITH J E, MANNUCCI P V. From creativity to innovation:the social network drivers of the four phases of the idea journey [J]. Academy of management review, 2017, 42(1):53 – 79.
[26] LIU T Y. Learning to rank for information retrieval[C]//Interna-

tional ACM SIGIR conference on research & development in information retrieval. Geneva: ACM Press. 2010:1-112.
[27] RADLINSKI F, JOACHIMS T. Query chains: learning to rank from implicit feedback[C]// ACM SIGKDD international conference on knowledge discovery & data mining. Chicago: ACM Press, 2005:239-248.
[28] RADLINSKI F, JOACHIMS T. Active exploration for learning rankings from clickthrough data[C]//ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM Press 2007: 570-579.
[29] HOSSEINI M, MOORE J, ALMALIKI M, et al. Wisdom of the crowd within enterprises: practices and challenges [J]. Computer networks, 2015, 17(15): 121-132.
[30] 熊李艳,陈晓霞,钟茂生,等. 基于PairWise排序学习算法研究综述[J].科学技术与工程,2017,17(21):184-190.
[31] JOACHIMS T. Training linear SVMs in linear time[C]// ACM SIGKDD international conference on knowledge discovery & data mining. Philadelphia:ACM Press, 2006:217-226.
[32] Sourceforge[EB/OL].[2020-05-08]. https://sourceforge.net/p/lemur/wiki/RankLib%20How%20to%20use/.

作者贡献说明:

易明:研究思路和论文大纲的提出,论文定稿;
张婷婷:数据采集与分析,论文初稿撰写;
李梓奇:数据分析优化,论文修改。

Research on the Ranking of Social Q&A Community Answers Based on Multidimensional Features

Yi Ming Zhang Tingting Li Ziqi
School of Information Management, Central China Normal University, Wuhan 430079

Abstract: [**Purpose/significance**] This paper studies the impact of multi-dimensional characteristics on Social Q&A Communities answer ranking, which can improve the service quality in Social Q&A Communities and optimize the user experience. [**Method/process**] This paper constructed a Social Q&A Communities answer ranking feature system from the answer feature, respondent feature and voter feature dimensions, and then we compared the applicability of 11 ranking learning algorithms based on deep learning, tree, neural network and support vector machine in Social Q&A Communities data set, and train random forest classification algorithm to get the importance of each feature. [**Result/conclusion**] The experimental results show that the sorting learning algorithm based on deep learning performs better than other sorting algorithms in NDCG@ k and MRR indexes, and the influence characteristics of voters are very important, followed by the content characteristics of the answers, and finally the professional characteristics of the respondents. From the two dimensions of increasing the diversity of the answer ranking method and improving the comprehensiveness of the answer ranking algorithm, we provide some suggestions for the optimization of community answer ranking.

Keywords: Social Q&A Community answer quality ranking learning algorithm deep learning algorithm